

Article

# Gráfico de control T2 Hotelling para variables cualitativas

Wilson Rojas-Preciado<sup>1,2</sup>, Mauricio J. Rojas-Campuzano<sup>3</sup>, Purificación Galindo-Villardón<sup>2</sup>, Omar Ruiz-Barzola<sup>3</sup>, , , ,

\* Correspondence: [wrojas@utmachala.edu.ec](mailto:wrojas@utmachala.edu.ec); Tel.: +593-992-83-3719

† Current address: Updated affiliation

‡ These authors contributed equally to this work.

Version November 20, 2021 submitted to Water



**Simple Summary:** A Simple summary goes here.

**Abstract:** Abstract

**Keywords:** keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

## 1. Introduction

Los gráficos de control constituyen una de las herramientas más importantes para definir límites y parámetros óptimos de los procesos, así como para controlar la calidad de los productos mediante la reducción de la variabilidad. El uso de gráficos de control facilita la evaluación del comportamiento de las variables del proceso y contribuye al logro de los objetivos planificados.

La variación de los procesos se entiende como la diversidad de resultados que genera un grupo de variables de un proceso, su monitoreo es un objetivo clave del control estadístico, por lo tanto, es necesario entender los tipos y motivos de la variabilidad. Para ello es preciso registrar de manera sistemática y adecuada diferentes variables del proceso que se desea controlar, como las propiedades de los insumos, las condiciones de operación de los equipos, las competencias del personal que maneja los procesos, además de las características de los productos, la satisfacción de los usuarios, el cumplimiento de requisitos, entre otras.

El pionero del control estadístico de procesos fue Walter Shewhart. Estableció las diferencias entre la variabilidad natural o común, presente en todos los procesos, y la provocada por causas asignables o especiales, que pueden llevarlos a un estado de fuera de control. Señaló que un proceso está en control estadístico cuando trabaja sólo con causas comunes de variación. Propuso los primeros gráficos de control para variables de tipo continuo y para variables de atributos [1].

El control estadístico de procesos mediante gráficos de control permitió a las organizaciones monitorear el comportamiento de una variable a la vez, no obstante, las organizaciones requirieron, con el pasar del tiempo, el análisis de varias características de calidad de forma simultánea, abriendo la puerta al control estadístico de procesos desde una perspectiva multivariante [2]. Para facilitar el control de calidad de procesos es común el uso de gráficas de control que recolectan abundante información en diversas variables de forma simultánea, su análisis permite caracterizar los diferentes tipos de variables que afectan la calidad y explican su comportamiento a lo largo del tiempo [3].

Hay una variedad de gráficos de control de procesos desde la perspectiva multivariante, entre los clásicos están el Gráfico  $T^2$  de Hotelling [4], el Multivariate Exponentially Weighted Moving –

MEWMA [5], el Multivariate Cumulative Sum Control Chart – MCUSUM [6]. Con el transcurso del tiempo se hicieron diversos aportes para mejorar el rendimiento de estos gráficos, entre los más destacados están Gráfico de control  $T^2$  con tamaños de muestra adaptables [7], Gráfico de control  $T^2$  con intervalos de muestreo variables [8], Gráfico de control  $T^2$  con líneas de advertencia dobles [9], Gráfico de control robusto [10], Gráficos de control basados en modelos de minería de datos para procesos multivariantes y autocorrelacionados [11], Gráficos de control de calidad multivariantes con dimensión variable [12], Gráfico de control para el coeficiente de variación multivariante [13].

Además de estos gráficos de control para entornos paramétricos, se desarrollaron otros para datos numéricos y cualitativos en entornos multivariantes no paramétricos, entre ellos el Gráfico de control multivariante basado en la distancia de Gower para una combinación de variables continuas y cualitativas [14], Gráfico de control multivariante basado en la combinación de PCA para características de calidad de atributos y variables [15], Gráfico de control multivariante no paramétrico basado en la ponderación de novedad sensible a la densidad para procesos no normales [16], Gráfico de control de deméritos con clustering difuso de c-medias [17], Gráfico de control basado en ACP que utiliza máquinas de vectores de soporte para distribuciones no normales multivariadas [18], Gráfico CUSUM no paramétrico para monitorear procesos multivariados correlacionados en serie [19], Gráfico de control multivariante basado en Kernel PCA para monitorear características de calidad de atributos y variables mixtas [20], Gráfico  $T^2$  basado en la combinación de PCA para datos continuos y cualitativos con detección de datos atípicos [21].

Como se puede observar, la literatura científica es abundante en lo referente a gráficos de control en entornos multivariantes paramétricos y no paramétricos para datos numéricos y, en los últimos años, para datos mixtos (numéricos y cualitativos), sin embargo, son pocas las publicaciones sobre gráficos de control multivariantes para datos cualitativos. En este campo las propuestas se han desarrollado alrededor del análisis de variables que siguen una distribución Poisson y el análisis de variables multinomiales.

La primera propuesta fue la de Holgate [22], quien presentó un trabajo sobre la distribución Poisson bivalente para variables correlacionadas. Este modelo fue tomado como insumo en las investigaciones de autores como Chiu and Kuo [23], Lee and Costa [24], Laungrungrong *et al.* [25], Epprecht *et al.* [26]. Otra propuesta destacada es la de Lu [27], quien desarrolló un gráfico de control tipo Shewhart para procesos multivariados con variables de atributos, cuando la característica de calidad asume valores binarios, que se denominó gráfico np multivariante (MNP). No obstante, hay escenarios en los que una clasificación dicotómica es insuficiente y se vuelve necesario acudir a niveles intermedios, en cuyo caso el análisis requiere el uso de distribuciones multinomiales.

En este contexto Mukhopadhyay [28] planteó un gráfico de control multivariante utilizando el estadístico D2 de Mahalanobis para atributos que siguen una distribución multinomial. Además, surgieron los gráficos de control multivariantes en procesos multinomiales bajo el enfoque difuso [29]; Taleb [30] introdujo gráficos de control para el monitoreo de procesos multivariados con datos lingüísticos multidimensionales, basados en dos procedimientos: la teoría de la probabilidad y la teoría difusa; Fernández *et al.* [31] presentaron un gráfico de control multivariante multinomial T2 con un enfoque difuso.

Otra propuesta interesante es la de Epprecht *et al.* [26], quienes presentaron una combinación lineal óptima de variables discretas, cuando siguen la distribución de Poisson, para el control de procesos estadísticos multivariados.

En el estudio de los procesos que se desarrollan en el entorno social-educativo y que explican el comportamiento de variables como el rendimiento académico, tasas de graduación o deserción, producción científica, porcentajes de matrícula de nuevo ingreso, entre otros, se maneja con mucha frecuencia variables cualitativas. No es que estén ausentes los datos cuantitativos, sino que, en las bases de datos que se utilizan para estos análisis, abundan las variables cualitativas nominales y ordinales sobre las de tipo numérico, algunos ejemplos de datos de los estudiantes son: sexo, lugar de procedencia, autodenominación étnica, grado académico de los padres, tipo de institución

educativa de procedencia (fiscal, particular, municipal); ejemplos de datos de las instituciones son: tipo de sostenimiento económico, jornada, modalidad, campo de estudio, niveles (tecnológico, grado y postgrado), tipo de infraestructura; ejemplos asociados a datos de los profesores son: titularidad, dedicación, grado académico, grado en el escalafón, discapacidad, entre otros.

López [32] señala que al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los métodos multivariantes de reducción de la dimensión tratan de eliminarla combinando muchas variables observadas para quedarse con pocas variables ficticias que, aunque no observadas, sean combinación de las reales y sinteticen la mayor parte de la información contenida en sus datos. En este caso se deberá tener en cuenta el tipo de variables que maneja. Si son variables cuantitativas las técnicas que le permiten este tratamiento pueden ser el Análisis de componentes principales [33,34], el Análisis factorial [35–37], mientras que, si se trata de variables cualitativas, es recomendable la aplicación de un Análisis de correspondencias múltiple, Análisis de homogeneidad o un Análisis de Escalamiento multidimensional.

### 1.1. Análisis de Correspondencias

El tratamiento multivariante de variables cualitativas requiere un proceso metodológico distinto, uno de los más representativos es el Análisis de Correspondencias [38]. Según [32], este análisis implica estudios de similaridad o disimilaridad entre categorías, se debe cuantificar la diferencia o distancia entre ellas sumando las diferencias cuadráticas relativas entre las frecuencias de las distribuciones de las variables analizadas, lo que conduce al concepto de la  $\chi^2$ . Así, el análisis de correspondencias puede considerarse como un análisis de componentes principales aplicado a variables cualitativas que, al no poder utilizar correlaciones, se basa en la distancia no euclídea de la  $\chi^2$ . En el enfoque francés del Análisis de Correspondencias, que se caracteriza por el énfasis en la geometría, el análisis de una tabla cruzada se llama análisis de correspondencia (CA) y el análisis de una colección de matrices indicadoras, se denomina análisis de correspondencia múltiple (MCA) [39]. En contextos anglosajones, el MCA es conocido como Análisis de Homogeneidad o Escalamiento Dual, especialmente en psicometría.

### 1.2. Análisis de Homogeneidad

El Análisis de Homogeneidad, Homogeneous Alternating Least Squares (HOMALS), es un modelo de la familia de modelos matemáticos del Escalamiento óptimo del sistema Gifi [40], el cual comprende una serie de técnicas exploratorias de análisis multivariado no lineal. Igual que el MCA, HOMALS se considera una forma de Análisis de Componentes Principales para datos cualitativos. El Análisis de Homogeneidad representa los objetos analizados mediante puntos en el modelo espacial, sus características más relevantes se presentan en las relaciones geométricas entre los puntos, para ello, es necesario la cuantificación de datos cualitativos [41]. El uso de variables cualitativas no es particularmente restrictivo, ya que una variable numérica continua se puede considerar como una variable cualitativa con un gran número de categorías. HOMALS se diferencia del el MCA en que éste utiliza la función de Descomposición de valores propios mientras que el Análisis de Homogeneidad utiliza Mínimos Cuadrados Alternos, lo que se conoce en la literatura como la Solución de Homals [39].

### 1.3. Escalamiento multidimensional

Otra de las técnicas multivariantes para el tratamiento de variables cualitativas es el escalamiento multidimensional (EMD) [42,43]. Se define como una técnica que elabora una representación gráfica que permite conocer la imagen que los individuos crean de un conjunto de objetos por posicionamiento de cada uno en relación con los demás, (mapa perceptual). El EMD trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos. Esto se logra asignando las observaciones a posiciones específicas en un espacio conceptual, normalmente de dos o tres dimensiones, de modo que las distancias entre los puntos en el espacio concuerden al máximo con las disimilaridades dadas. Las

dimensiones de este espacio conceptual se pueden interpretar para favorecer la comprensión de los datos, inclusive si son valoraciones subjetivas de disimilaridad entre objetos o conceptos. Dado que el EMD permite calcular las distancias a partir de los datos multivariados, si las variables se han medido objetivamente, se lo puede utilizar como técnica de reducción de datos [32].

Las variables que utiliza esta técnica pueden ser métricas o no métricas. El escalamiento multidimensional las transforma en distancias entre los objetos en un espacio de dimensiones múltiples, de modo que objetos que aparecen situados más próximos entre sí son percibidos como más similares por los sujetos.

## 2. Materials and Methods

### 2.1. Notación

La tabla 1 contiene elementos, representación y ejemplos de la manera como se presentan los elementos algebraicos abordados en la metodología.

Elementos	Representación	Ejemplo
Escalares	Letras en minúscula.	$v, \lambda$
Vectores	Letras en minúscula y en negrita.	$\mathbf{v}, \mathbf{u}$
Matrices	Letras en mayúscula y en negrita.	$\mathbf{V}, \mathbf{X}$
Matrices de tres vías (Cubos de datos)	Letras con doble trazo en mayúscula.	$\mathbb{C}, \mathbb{X}$

**Table 1.** Elementos algebraicos

A lo largo del artículo se utilizarán letras para hacer referencia a parámetros necesarios, se los enuncia a continuación en la tabla 2:

Letra	Significado	Especificación
$p$	Número de dimensiones	
$K$	Número total de tablas (Especifica la profundidad del cubo de datos)	
$k$	Índice de tabla	$k=1,2,\dots,K$
$T$	Índice de matriz transpuesta	$\mathbf{X}^T$
$n$	Tamaño muestral de las $K$ tablas	

**Table 2.** Notación

### 2.2. Análisis de Correspondencia Múltiple (MCA)

El análisis de correspondencias múltiples (MCA) es una generalización del análisis de correspondencias simple o binario, donde se incluyen más variables cualitativas. Se obtiene al realizar el análisis de correspondencias simple a una tabla disyuntiva completa, conocida como la tabla de Burt.

Se tiene una matriz de datos con  $p$  variables cualitativas, cada una con  $h$  categorías ( $h > 1$ ). En el ejemplo que se desarrolla para esta investigación, se dispone de una base de datos (*Datak10Contaminated*) constituida por 10 tablas, cada una tiene 10 variables y cada variable, 3 categorías (Alto, Medio y Bajo).

Esta matriz es equivalente a la matriz disyuntiva  $Z$ , que desglosa las variables en cada una de sus modalidades y se registra la ocurrencia de eventos de forma binaria.

La tabla de Burt viene dada por:

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z} \quad (1)$$

$V_1$	$V_2$	$\cdots$	$V_p$
Alto	Medio	$\cdots$	Medio
Medio	Bajo	$\cdots$	Alto
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Bajo	Alto	$\cdots$	Bajo

**Table 3.** Matriz inicial

$V_1$ : Alto	$V_1$ : Medio	$V_1$ : Bajo	$V_2$ : Alto	$V_2$ : Medio	$V_2$ : Bajo	$\cdots$	$V_p$ : Alto	$V_p$ : Medio	$V_p$ : Bajo
1	0	0	1	0	0	$\cdots$	0	1	0
0	1	0	0	1	0	$\cdots$	1	0	0
0	0	1	0	0	1	$\cdots$	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
1	0	0	1	0	0	$\cdots$	1	0	0

**Table 4.** Matriz disyuntiva Z

La construcción de la matriz de Burt se da por la superposición de tablas. En las tablas ubicadas en la diagonal se encuentran matrices diagonales que contienen las frecuencias marginales de cada una de las variables. Fuera de la diagonal de la matriz de Burt se encuentran las tablas cruzadas por pares de variables.

Para realizar el análisis de correspondencias múltiples se parte de la matriz de Burt, obtenida con la ecuación 1. Esta matriz está formada por las frecuencias absolutas, éstas se transforman en frecuencias relativas, dividiendo los valores de la matriz por la frecuencia total, dando lugar a una nueva matriz que se denominará **P**.

	$V_1$ : Alto	$V_1$ : Medio	$V_1$ : Bajo	$V_2$ : Alto	$V_2$ : Medio	$V_2$ : Bajo	$\cdots$	$V_p$ : Alto	$V_p$ : Medio	$V_p$ : Bajo
$V_1$ : Alto	$b_{1,1}$	0	0	$b_{1,4}$	$b_{1,5}$	$b_{1,6}$	$\cdots$	$b_{1,3p-2}$	$b_{1,3p-1}$	$b_{1,3p}$
$V_1$ : Medio	0	$b_{2,2}$	0	$b_{2,4}$	$b_{2,5}$	$b_{2,6}$	$\cdots$	$b_{2,3p-2}$	$b_{2,3p-1}$	$b_{2,3p}$
$V_1$ : Bajo	0	0	$b_{3,3}$	$b_{3,4}$	$b_{3,5}$	$b_{3,6}$	$\cdots$	$b_{3,3p-2}$	$b_{3,3p-1}$	$b_{3,3p}$
$V_2$ : Alto	$b_{4,1}$	$b_{4,2}$	$b_{4,3}$	$b_{4,4}$	0	0	$\cdots$	$b_{4,3p-2}$	$b_{4,3p-1}$	$b_{4,3p}$
$V_2$ : Medio	$b_{5,1}$	$b_{5,2}$	$b_{5,3}$	0	$b_{5,5}$	0	$\cdots$	$b_{5,3p-2}$	$b_{5,3p-1}$	$b_{5,3p}$
$V_2$ : Bajo	$b_{6,1}$	$b_{6,2}$	$b_{6,3}$	0	0	$b_{6,6}$	$\cdots$	$b_{6,3p-2}$	$b_{6,3p-1}$	$b_{6,3p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$V_p$ : Alto	$b_{3p-2,1}$	$b_{3p-2,2}$	$b_{3p-2,3}$	$b_{3p-2,4}$	$b_{3p-2,5}$	$b_{3p-2,6}$	$\cdots$	$b_{3p-2,3p-2}$	0	0
$V_p$ : Medio	$b_{3p-1,1}$	$b_{3p-1,2}$	$b_{3p-1,3}$	$b_{3p-1,4}$	$b_{3p-1,5}$	$b_{3p-1,6}$	$\cdots$	0	$b_{3p-1,3p-1}$	0
$V_p$ : Bajo	$b_{3p,1}$	$b_{3p,2}$	$b_{3p,3}$	$b_{3p,4}$	$b_{3p,5}$	$b_{3p,6}$	$\cdots$	0	0	$b_{3p,3p}$

**Table 5.** P: Tabla de contingencia de Burt en frecuencias relativas

Se obtienen las marginales de las filas (*mf*) y de las columnas (*mc*) de la matriz **P** (Tabla 5). A estos vectores se los conoce también como *Masas de fila y columna*, respectivamente.

$V_1$ : Alto	$V_1$ : Medio	$V_1$ : Bajo	$V_2$ : Alto	$V_2$ : Medio	$V_2$ : Bajo	$\cdots$	$V_p$ : Alto	$V_p$ : Medio	$V_p$ : Bajo
$b_{\bullet,1}$	$b_{\bullet,2}$	$b_{\bullet,3}$	$b_{\bullet,4}$	$b_{\bullet,5}$	$b_{\bullet,6}$	$\cdots$	$b_{\bullet,3p-2}$	$b_{\bullet,3p-1}$	$b_{\bullet,3p}$

**Table 6.** Frecuencias marginales de las filas. (mf)

$V_1$	:	$V_1$	:	$V_1$	:	$V_2$	:	$V_2$	:	$V_2$	:	$\dots$	$V_p$	:	$V_p$	:	$V_p$	:
<i>Alto</i>		<i>Medio</i>		<i>Bajo</i>		<i>Alto</i>		<i>Medio</i>		<i>Bajo</i>			<i>Alto</i>		<i>Medio</i>		<i>Bajo</i>	
$b_{\bullet,1}$		$b_{\bullet,2}$		$b_{\bullet,3}$		$b_{\bullet,4}$		$b_{\bullet,5}$		$b_{\bullet,6}$		$\dots$	$b_{\bullet,3p-2}$		$b_{\bullet,3p-1}$		$b_{\bullet,3p}$	

**Table 7.** Frecuencias marginales de las columnas. (mc)

Se obtiene la matriz de residuos estandarizados  $S$ .

$$S = D_{\text{fila}}^{-\frac{1}{2}}(P - mf \ mc')D_{\text{columna}}^{-\frac{1}{2}} \quad (2)$$

donde:

- $D_{\text{fila}}$  es una matriz diagonal que contiene las masas de las filas.
- $D_{\text{columna}}$  es una matriz diagonal que contiene las masas de las columnas

Se aplica descomposición singular (SVD) a la matriz  $S$  (Ecuación 2):

$$S = UDV' \quad (3)$$

donde:

- $U$  y  $V$  son matrices ortogonales.
- $D$  es una matriz diagonal que contiene los valores singulares.

Para encontrar las coordenadas estandarizadas se aplica lo siguiente:

$$X = D_{\text{fila}}^{-\frac{1}{2}}U \quad (4)$$

$$Y = D_{\text{columna}}^{-\frac{1}{2}}V \quad (5)$$

Para los fines necesarios, se utilizará las coordenadas de las columnas (Tabla 8).

	$Dim_1$	$Dim_2$	$\dots$	$Dim_{3p}$
$V_1 : Alto$	$v_1d_{1alto}$	$v_1d_{1alto}$	$\dots$	$v_1d_{palto}$
$V_1 : Medio$	$v_1d_{1medio}$	$v_1d_{1medio}$	$\dots$	$v_1d_{pmedio}$
$V_1 : Bajo$	$v_1d_{1bajo}$	$v_1d_{1bajo}$	$\dots$	$v_1d_{pbajo}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$V_p : Bajo$	$v_pd_{1bajo}$	$v_pd_{1bajo}$	$\dots$	$v_pd_{pbajo}$

**Table 8.** Coordenadas estandarizadas de las columnas.

### 2.3. Generalización a K tablas

Si se tienen K tablas, con la misma estructura de la tabla 3, como se visualiza en la figura 1, se aborda el enfoque del análisis factorial múltiple (MFA). Escofier and Pagès [44] indica que el MFA utiliza análisis de correspondencia múltiple cuando se trata de variables cualitativas. El procedimiento implica la realización de un MCA por cada tabla y dividirlo por su primer valor propio con la finalidad de obtener K grupos normalizados. Posteriormente se consideran todas las tablas y se realiza un MCA global.

La generalización a K tablas del procedimiento del MCA, se presenta en la Figura 2

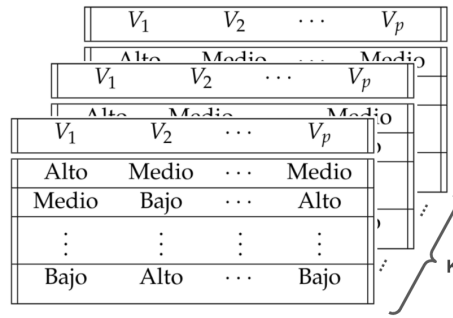


Figure 1. K tablas con el formato inicial.

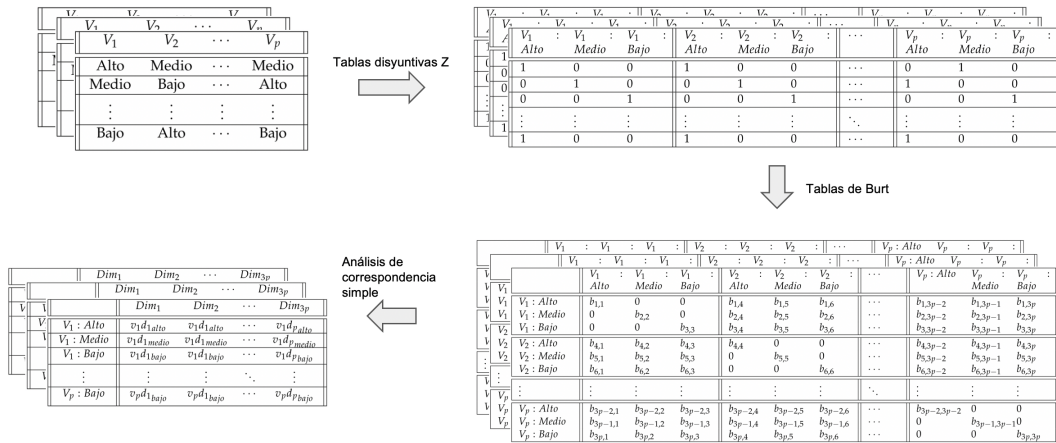


Figure 2. Procedimiento del MCA para K tablas

Se llama  $C$  a cada tabla de coordenadas. Con la finalidad de detectar la magnitud de las variables latentes, su aporte neto a las variables, se trata la matriz  $C$  con valor absoluto.

#### 2.4. Aporte del Análisis Factorial Múltiple (MFA)

Una vez que se tienen las coordenadas de las columnas, se procede a realizar la normalización, característica del procedimiento MFA.

Sea  $\lambda_1^k$  el primer valor propio obtenido de la descomposición singular de la  $k$ -ésima tabla  $C$ . Se normaliza la tabla multiplicándola por  $1/\lambda_1^k$ . Con esto se obtiene la tabla  $C'$ , que corresponde a la tabla de coordenadas normalizadas.

Individualmente, para el caso de la matriz  $k$ , se tendría la siguiente expresión.

$$C'_k = \frac{1}{\lambda_1^k} C_k \quad (6)$$

Aglomerando las matrices normalizadas  $C'$  en una sola, se tiene la matriz  $C'$ . Esta contiene todos los elementos de las  $k$  tablas.

$$C' = [C'_1 | C'_2 | \dots | C'_k]^T \quad (7)$$

La normalización que realiza el MFA se encarga de ponderar las  $k$  tablas, con el objetivo de evitar alguna descompensación al momento de realizar el análisis conjunto de las tablas.

#### 2.5. Gráfico de control

Para definir el gráfico de control  $T^2$  Hotelling se deben tomar las siguientes consideraciones:



- La tabla  $\mathbb{C}'$  (Ecuación 7) se denomina Consenso, sirve como referente para el escenario *bajo control*, y de la cual se obtiene  $\mu_0$  y  $\mathbf{S}_0$ .
- Cada matriz  $\mathbf{C}'_k$  tiene el mismo número de filas ( $n$ ) y columnas ( $p$ ) (individuos y variables).
- El vector de medias  $\bar{\mathbf{z}}_k$  está atado a la tabla  $\mathbf{C}'_k$ , es decir, el gráfico de control estará en función de las diferencias entre las matrices  $\mathbf{C}'_k$  y la matriz consenso  $\mathbb{C}'$ .
- Las matrices  $\mathbf{C}'_k$  siguen una distribución normal multivariante con vector de medias  $\mu_k$  y matriz de covarianzas  $\mathbf{S}_k$ .

Con esto se obtiene el estadístico  $T^2$ :

$$T^2 = n(\mu_k - \mu_0)' \Sigma_0^{-1} (\mu_k - \mu_0) \quad (8)$$

Se sabe que, bajo control, el  $T^2$  se distribuye como una Chi-cuadrado con  $p$  grados de libertad  $\chi_p^2$ . En este caso se puede aplicar este principio, ya que se utiliza la matriz consenso ( $\mathbb{C}'$ ), que representa al escenario bajo control.

Dado que este gráfico de control está basado en distancias de Mahalanobis ponderadas, sólo tiene límite de control superior. Este viene dado por la ecuación 9

$$UCL = \chi_{\alpha, p}^2 \quad (9)$$

donde  $p$  es el número de dimensiones y  $\alpha$  es la significancia predeterminada considerando  $p$ .

## 2.6. Tabla posterior

Con la finalidad de detectar las potenciales categorías responsables de que un punto en el gráfico  $T^2$  de Hotelling para variables cualitativas se encuentre fuera de control, se propone una tabla que presenta las anomalías de cada categoría en cada variable, comparando las masas de columna de la tabla  $k$  y las masas de columna de la tabla consenso por medio de distancias  $\chi^2$  que proporcionan un valor  $p$ , aportando a la interpretación.

## 3. Complemento computacional

Para facilitar la difusión y aplicación del método propuesto, se ha desarrollado un paquete reproducible en R. El paquete **T2Qv** utiliza la metodología expuesta en este artículo y la lleva a un entorno práctico, permite visualizar los resultados de forma plana o interactiva, además, presenta un panel Shiny que contiene todas las funciones individuales en un mismo espacio.

### 3.1. Disponibilidad

El paquete está disponible en GitHub, la descarga se la puede realizar de la siguiente forma:

```
install.packages("devtools")
devtools::install_github("JavierRojasC/T2Qv")
```

### 3.2. El paquete: T2Qv

Las funciones que contiene el paquete y su descripción se enuncian en la tabla 9.

## 4. Resultados

Con la intención de probar la metodología propuesta en el gráfico de control  $T^2$  de Hotelling para variables cualitativas, se hizo un análisis con datos simulados y otro con datos reales aplicados al contexto de la educación superior. Los resultados se obtienen de la aplicación del paquete T2Qv.



```

Package: T2Qv
Type: Package
Title: Control Qualitative Variables
Version: 0.1.0
Authors@R: c(person("Wilson", "Rojas-Preciado", role = c("aut", "cre"),
  email = "wrojas@utmachala.edu.ec"),
  person("Mauricio", "Rojas-Campuzano", role = c("aut", "ctb"),
  email = "mauroja@espol.edu.ec"),
  person("Purificación", "Galindo-Villardón", role = c("aut", "ctb"),
  email = "oruiz@espol.edu.ec"),
  person("Omar", "Ruiz-Barzola", role = c("aut", "ctb"),
  email = "oruiz@espol.edu.ec"))
Maintainer: Wilson Rojas-Preciado <wrojas@utmachala.edu.ec>
Description: Covers k-table control analysis using multivariate control charts for qualitative variables using
fundamentals of multiple correspondence analysis and multiple factor analysis. The graphs can be shown in a
flat or interactive way, in the same way all the outputs can be shown in an interactive shiny panel.
License: MIT + file LICENSE
Encoding: UTF-8
LazyData: true
RoxygenNote: 7.1.1
Depends: R (>= 2.10)
Imports: shiny, shinydashboardPlus, shinydashboard, shinycssloaders,
  dplyr, ca, highcharter, stringr, tables, htmltools (>= 0.5.1.1)
Suggests: testthat (>= 3.0.0)
Config/testthat/edition: 3
Author: Wilson Rojas-Preciado [aut, cre],
  Mauricio Rojas-Campuzano [aut, ctb],
  Purificación Galindo-Villardón [aut, ctb],
  Omar Ruiz-Barzola [aut, ctb]
Built: R 4.0.2; ; 2021-10-14 23:56:56 UTC; unix

```

**Figure 3.** Documentación del paquete T2Qv

Función	Descripción
T2 qualitative	Multivariate control chart T2 Hotelling applicable for qualitative variables.
MCAconsensus	Multiple correspondence analysis applied to a consensus table.
MCApoint	Multiple correspondence analysis applied to a specific table.
ChiSq variable	Contains Chi square distance between the column masses of the table specified in PointTable and the consensus table. It allows to identify which mode is responsible for the anomaly in the table in which it is located.
Full Panel	A shiny panel complete with the multivariate control chart for qualitative variables, the two MCA charts and the modality distance table. Within the dashboard, arguments such as type I error and dimensionality can be modified.

**Table 9.** Funciones del paquete T2Qv

#### 4.1. Resultados con datos simulados

Para este estudio se generó una base de datos simulados, a la que se denominó *Datak10Contaminated*. Consta de 10 tablas, cada una de ellas está constituida por 100 filas (observaciones) y 11 columnas, de las cuales, las 10 primeras corresponden a las variables analizadas ( $V_1, V_2, \dots, V_{10}$ ) mientras que, la columna 11, denominada *GroupLetter*, contiene el factor de clasificación de los grupos. Para su identificación, las tablas han sido denominadas con las letras del alfabeto, desde la *a* hasta la *j*. La tabla *j* tiene una distribución distinta de la que tienen las otras nueve. Los datos se expresan en tres niveles: alto, medio y bajo. La tabla ?? presenta las 10 primeras filas de la base de datos *Datak10Contaminated*.

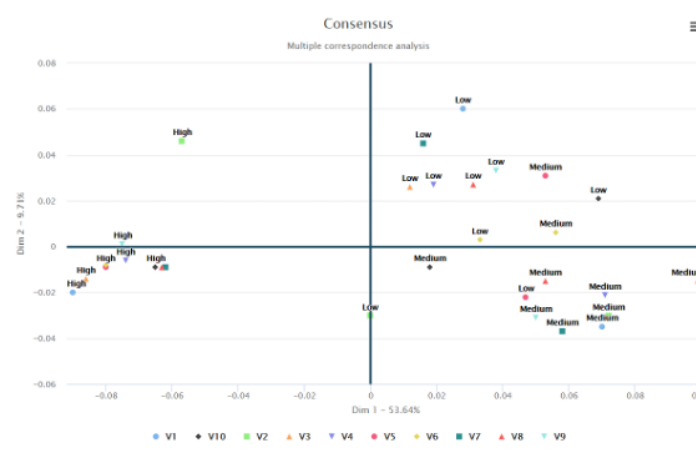
Para facilitar análisis se creó un paquete al que se denominó *T2Qv*, herramienta diseñada en el software estadístico R y R Studio. *T2Qv* realiza el análisis de control de *k* tablas utilizando gráficos de control multivariantes para variables cualitativas, utilizando los fundamentos del análisis de correspondencia múltiple y el análisis de factores múltiples. Los gráficos se pueden mostrar de forma plana o interactiva, de la misma manera todas las salidas se pueden mostrar en un panel interactivo de Shiny. El primer resultado es el gráfico del Análisis de Correspondencias Múltiple (MCA) aplicado a la tabla consenso (Figura 5). Esta tabla ha sido tomada como referente, como escenario en control para el análisis posterior de las tablas que sean reportadas como puntos fuera de control en el gráfico T2 de Hotelling. El MCA reporta una inercia total del 63.3%, la dimensión 1 representa al 53.6% de la información, mientras que la dimensión 2, al 9.7%. Los puntos del gráfico representan a las

Tabla n. Sección de la base de datos Datak10Contaminated

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	GroupLetter
Low	Medium	Medium	High	High	High	Low	Medium	Medium	Medium	a
Low	Low	High	Low	Medium	High	High	High	Low	High	a
High	Medium	High	Low	High	Medium	Medium	High	Medium	Low	a
Medium	Medium	Low	High	Low	Medium	High	Low	Low	High	a
Low	Low	Low	High	Low	High	High	High	Medium	Medium	a
High	High	Medium	Low	High	Low	Medium	Medium	High	Low	a
High	High	Low	Low	Low	Medium	High	Medium	Medium	High	a
Medium	Medium	High	Medium	Medium	High	Medium	High	High	High	a
Low	Low	Low	Medium	High	Medium	Low	Medium	Low	Low	a
Medium	Medium	Medium	High	Low	Medium	High	Low	High	Medium	a

**Figure 4.** Sección de la base de datos Datak10Contaminated.

observaciones de cada una de las 10 variables en sus tres niveles: alto, medio y bajo. En esta figura, todas las observaciones que corresponden al nivel alto se ubican a la izquierda en el eje de las X; de las 10 observaciones correspondientes al nivel medio, 8 se situaron en el cuarto cuadrante y las dos restantes en el cuadrante 1, es decir, todas las observaciones de este nivel estuvieron a la derecha en el eje de las X. Finalmente, de los 10 puntos que representan al nivel bajo, 8 están ubicados en el cuadrante 1.

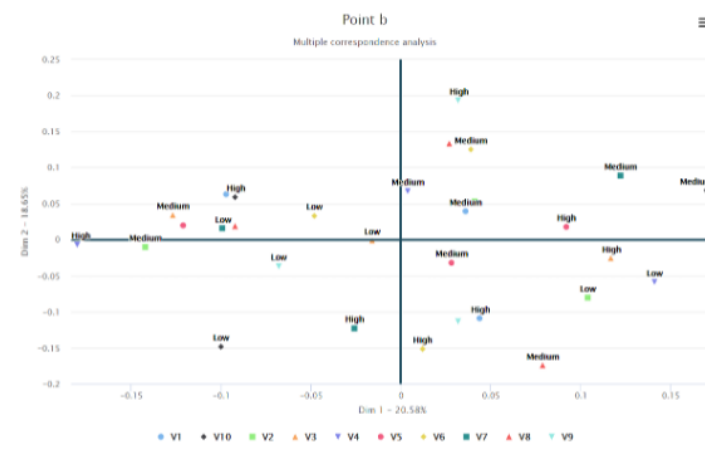
**Figure 5.** Análisis de correspondencias múltiple aplicado a la tabla consenso.

Otro resultado es el Análisis de Correspondencias Múltiple aplicado a una tabla específica. En este punto, uno de los argumentos que se debe tener en cuenta es la selección de la tabla de la que se realizará el análisis.

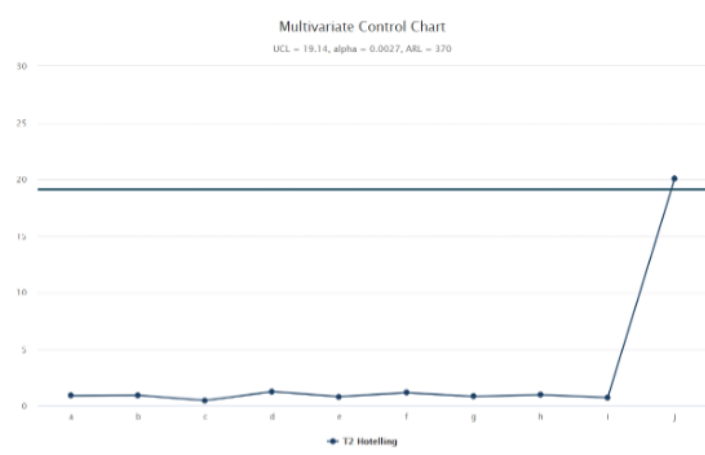
La figura 6 representa el gráfico del MCA de la tabla *b*. Este gráfico, en sus dos dimensiones, representa al 39.3% de la información. Es notorio que las observaciones en sus niveles alto, medio y bajo están distribuidas de forma aleatoria en todos los cuadrantes del del gráfico, no se puede precisar un patrón específico de agrupación. Esto mismo se puede decir de los puntos representados en cualquiera de las otras tablas, exceptuando la tabla *j*, que fue diseñada con una distribución diferente. No obstante, el uso del MCA de las figuras 6 y 5 todavía no permite detectar si el proceso está o no en control. La identificación de puntos fuera de control se puede realizar mediante la representación gráfica del estadístico T2 de Hotelling, como se observa en la figura 7.

La figura 7 presenta un gráfico de control elaborado con el estadístico T2 de Hotelling, aplicado a la detección de anomalías en cualquiera de las *k* tablas analizadas. Cada una de las tablas está representada por los puntos en el gráfico. Se observa una línea horizontal que representa al límite de control superior (UCL). El límite de control inferior (LCL) es igual a cero.

Dado que el análisis de sensibilidad determinó que este gráfico de control tiene un mejor rendimiento cuando trabaja con un número alto de dimensiones, se ha recomendado que este sea  $p-1$ , donde  $p$  es el



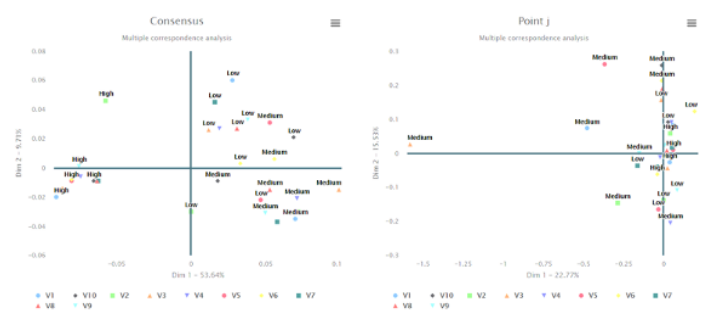
**Figure 6.** Análisis de correspondencias múltiple aplicado a la tabla b.



**Figure 7.** Gráfico de control multivariante T2 Hotelling aplicable a variables cualitativas.

número de dimensiones inicial, que es equivalente a la cantidad de variables de la base de datos, sin contar a la variable GroupLetter que sólo sirve como factor de clasificación de las tablas.

Se observa que el punto que representa a la tabla *j* se ubica por encima del límite de control superior, lo que quiere decir que se lo ha identificado como un valor fuera de control. Por consiguiente, es necesario analizar con detenimiento qué está pasando con los datos de la tabla reportada, comparándolos con los de la tabla consenso, a fin de identificar las causas de la variación y tomar las acciones pertinentes. Para hacer un análisis del punto fuera de control se realiza un gráfico del MCA de la tabla *j* y se lo compara con el gráfico similar de la tabla consenso, como se presenta en la figura 8.



**Figure 8.** Comparación de los gráficos del MCA aplicado a la tabla consenso y la tabla *j*.

La figura 8 presenta la distribución de las observaciones de las tablas consenso y  $j$  mediante gráficos del MCA. El gráfico de la tabla consenso, que sirve de referente en control, ya se analizó en la figura 5; el de la tabla  $j$  muestra una tendencia de los puntos que con valores medios a ubicarse al lado izquierdo, bastante alejados de los demás que confluyen hacia el centro del eje de las X. Especial atención merece la variable 3, que registra una observación para el nivel medio con el valor más alejado del grupo.

Al comparar los gráficos es obvio que la distribución de los datos en la tabla  $j$  es diferente de las distribuciones de las demás tablas, y en especial, es diferente de la distribución de la tabla consenso, lo que explica por qué el punto  $j$  ha sido identificado como fuera de control.

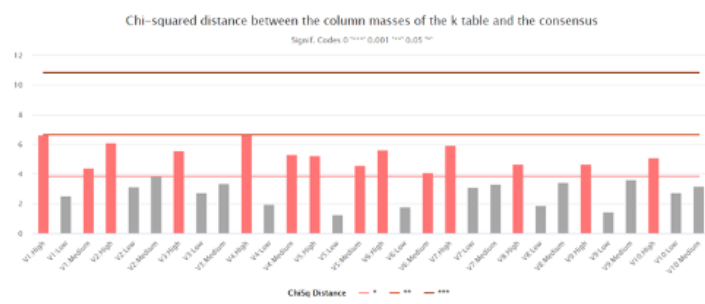
Para profundizar en el análisis se calcula las distancias Chi-cuadrado entre las masas de las columnas de la tabla reportada como fuera de control y la tabla consenso, tomada como referente. Estas distancias permiten identificar el comportamiento de las variables que inciden en el desplazamiento de la media del proceso que finalmente pueden llevarlo a un estado fuera de control (tabla 9).

La tabla 9 contiene los datos de cada una de las variables de la tabla  $j$  con sus tres niveles (alto, medio y bajo). La columna 3 muestra el p-valor para cada observación, de esto depende el número de asteriscos de la columna 4 que indica el nivel de significancia estadística. Así, si el p-valor es inferior a 0.05, la observación se reporta como significancia estadística y va asociada a un asterisco; si el p-valor es menor que 0.01, se entiende que hay alta significación estadística y se registran dos asteriscos; si el p-valor es menor que 0.001, la significancia estadística es muy alta y se reportan tres asteriscos; caso contrario, no se reporta significancia y la observación no lleva asteriscos.

Variable	Chi.Squared	val.p	Signif
V1:High	6.62953	0.01003	*
V1:Low	2.51447	0.11281	
V1:Medium	4.40573	0.03582	*
V2:High	6.10216	0.01350	*
V2:Low	3.15682	0.07561	
V2:Medium	3.81899	0.05067	
V3:High	5.58957	0.01807	*
V3:Low	2.73051	0.09845	
V3:Medium	3.37596	0.06615	
V4:High	6.61362	0.01012	*
V4:Low	1.95916	0.16160	
V4:Medium	5.33225	0.02093	*
V5:High	5.23785	0.02210	*
V5:Low	1.24566	0.26438	
V5:Medium	4.56461	0.03264	*
V6:High	5.64217	0.01753	*
V6:Low	1.81050	0.17845	
V6:Medium	4.11597	0.04248	*
V7:High	5.94284	0.01478	*
V7:Low	3.10801	0.07791	
V7:Medium	3.32453	0.06825	
V8:High	4.65021	0.03105	*
V8:Low	1.88624	0.16963	
V8:Medium	3.45642	0.06301	
V9:High	4.67660	0.03058	*
V9:Low	1.46059	0.22684	
V9:Medium	3.60793	0.05750	
V10:High	5.10688	0.02383	*
V10:Low	2.76592	0.09629	
V10:Medium	3.19604	0.07382	

**Figure 9.** Distancia  $\chi^2$  entre las masas de la tabla consenso y las k tablas, Datak10Contaminated

De las 30 observaciones que tiene la tabla  $j$ , se presentan 14 casos de p-valores menores que 0.05, es decir, reportan significancia estadística (un asterisco), de los cuales, 10 se atribuyen a las categorías altas de las variables cualitativas y cuatro a los niveles medios. El comportamiento de estas variables en la tabla  $j$ , que obedece a una distribución diferente a la de las demás tablas, provoca el desplazamiento de la media del proceso que, al final, lo lleva a un estado fuera de control. Otra manera de visualizar esta información es a través de un gráfico de barras (figura 10).



**Figure 10.** . Distancia  $\chi^2$  entre las masas de la tabla consenso y las k tablas, Datak10Contaminated.

En el gráfico de barras de la figura 10, se observa tres líneas horizontales que corresponden a los límites asociados a los niveles de significancia estadística, la más baja representa al p-valor inferior a 0.05 (un asterisco), la línea del medio, al p-valor inferior a 0.01 (dos asteriscos); y, la línea más alta representa al p-valor inferior a 0.001 (tres asteriscos). Las barras que representan valores sin significancia estadística no sobrepasan ninguna de las líneas y se pintan de color gris, mientras que, las que sí denotan significancia estadística adquieren el color de la línea más alta que sobrepasan.

#### 4.2. Resultados con datos aplicados al contexto de la educación superior

En este ejemplo se utiliza una base de datos denominada *Estudiantes 2019\_2020*, tomada de reportes que la Universidad Técnica de Machala (UTMACH) cargó en la plataforma del Sistema Integral de Información de la Educación Superior (SIIES), correspondiente a cuatro periodos académicos consecutivos. La base de datos *Estudiantes\_2019\_2020* contiene 43191 observaciones y 17 variables cualitativas referidas los estudiantes de las 30 carreras vigentes en sus 5 facultades.

Las variables registradas en la base de datos, con sus respectivas categorías son las siguientes:

- Periodo académico, esta es la variable que sirve como clasificador, hace referencia a 4 periodos de estudio (semestres): 2019-1, 2019-2, 2020-1 y 2020-2.

- Facultad, que tiene 5 categorías: Facultad de Ciencias Agropecuarias (FCA), Facultad de Ciencias Empresariales (FCE), Facultad de Ciencias Químicas y de la Salud (FCQS), Facultad de Ciencias Sociales (FCS) y Facultad de Ingeniería Civil (FIC).
- Carrera, variable que contiene el nombre de las 30 carreras vigentes en la UTMACH, cada una de ellas es una categoría y se asocia a alguna de las 5 facultades. En la FCA: Acuicultura, Economía Agropecuaria, Agronomía y Medicina Veterinaria; en la FCE: Administración de Empresas, Turismo, Mercadotecnia, Contabilidad y auditoría, Comercio internacional y Economía; en la FCQS: Medicina, Enfermería, Bioquímica y Farmacia, Alimentos, Ing. Química; en la FCS: Artes plásticas, Pedagogía de la Actividad Física y Deporte, Pedagogía de las Ciencias Experimentales, Educación Básica, Educación inicial, Pedagogía de los Idiomas Nacionales y Extranjeros, Psicología Clínica, Psicopedagogía, Comunicación, Derecho, Gestión Ambiental, Sociología, Trabajo Social; y en la FIC: Ingeniería Civil y Tecnología de la Información.
- Sexo, con sus dos clases: hombre y mujer.
- Grupo edad, que clasifica a los estudiantes en 5 grupos según su edad en años: Menores que 18, de 18 a 30, de 31 a 45, de 46 a 60 y Mayores a 60.
- Discapacidad, cuyas clases son: Intelectual, Auditiva, Física Motora, Visual, Lenguaje y Ninguna.

- Etnia, con sus tipos: Mestizo, Montubio, Negro, Blanco, Indígena, Mulato, Afroecuatoriano, Otro, No registra.
- Zona residencial, Urbana y Rural.
- Nivel de formación del padre: Centro de alfabetización, Educación Básica incompleta, Educación Básica, Bachillerato, Superior tecnológica incompleta, Superior tecnológica, Superior universitaria, Superior universitaria incompleta, Diplomado, Especialidad, Postgrado Maestría o Especialización en áreas de Salud, Postgrado Ph.D., Ninguno y No sabe, no registra.
- Nivel de formación de la madre: Centro de alfabetización, Educación Básica incompleta, Educación Básica, Bachillerato, Superior tecnológica incompleta, Superior tecnológica, Superior universitaria, Superior universitaria incompleta, Diplomado, Especialidad, Postgrado Maestría o Especialización en áreas de Salud, Postgrado Ph.D., Ninguno y No sabe, no registra.
- Número de miembros del hogar, con sus tres clases: Hasta 3, 4 y 5 o más.
- Tipo colegio: Fiscal, Particular, Fiscomisional, Extranjero, Municipal y No registra.
- Ingreso total hogar: Rango 1, Rango 2, Rango 3, Rango 4, Rango 5, Rango 6, Rango 7, Rango 8, Rango 9 y Rango 10.
- Origen de recursos estudios: Padres tutores, Hermanos y familiares, Pareja sentimental, Recursos propios, Beca estudio, Crédito educativo y No registra.
- Segunda matrícula: Sí y No.
- Tercera matrícula: Sí y No.
- Terminó periodo: Sí y No.

#### 4.2.1. Análisis de Correspondencias Múltiple de la tabla Consenso

El gráfico del Análisis de Correspondencias Múltiple que se realiza a la tabla consenso (Figura 5) es el escenario bajo control que se utilizará para el análisis de las tablas que luego se registren como puntos fuera de control en el gráfico T2 de Hotelling. El MCA reporta una inercia total del 30.31%. Los puntos del gráfico representan a las observaciones de cada una de las 16 variables en sus distintos niveles. La variable Periodo académico sirve como elemento clasificador, por eso sus observaciones no aparecen aquí.

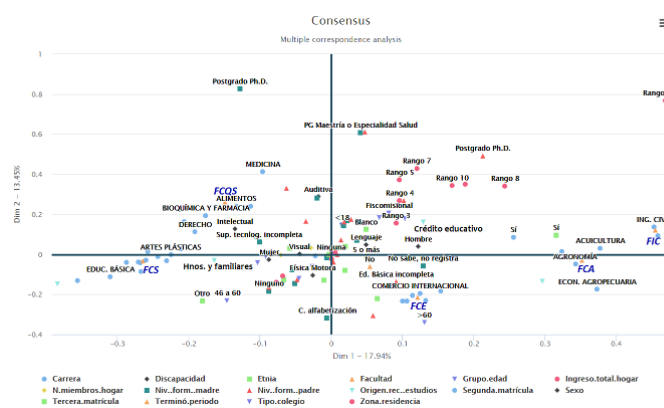


Figure 11. Gráfico de MCA de la tabla consenso, Estudiantes 2019-2020.

Se observa cómo las carreras se agrupan alrededor de sus respectivas facultades; la FIC y la FCA se muestran similares entre sí y ubicadas al lado derecho, en el plano que corresponde a la dimensión 1, mientras que, al otro extremo está la FCS. Por otra parte, las similitudes y diferencias entre las otras dos facultades giran en torno a los ejes de las dos dimensiones, la FCE ubicada en el cuadrante 4 y la FCQS, en el 2.

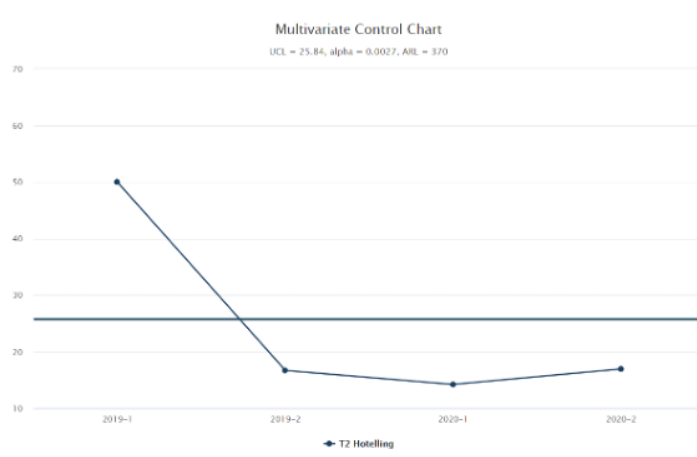
La variable Sexo es una de las que más incide en la ubicación de los puntos alrededor de la dimensión 1. El número de estudiantes varones es mayor que el de las mujeres en las carreras de la FCA y FIC, por otra parte, el número de mujeres es mayor que el de hombres en las carreras de la FCS y FCQS; en la FCE parece no haber marcada diferencia en la proporción de hombres y mujeres. Las variables Segunda

matrícula y Tercera matrícula dan cuenta de que es muy frecuente que los estudiantes aprueben sus asignaturas en su primera matrícula, sin repetir; se observa también que la segunda y tercera matrícula ocurren con mayor frecuencia en la FCA y la FIC, especialmente en ésta, lo que podría estar asociado al grado de dificultad propio de las asignaturas que allí se estudia, a procesos con mayor rigor académico y hasta a insuficiencias en los procesos de enseñanza – aprendizaje. Al otro extremo está la FCS, en la que no es usual que ocurran segundas o terceras matrículas.

La variable Ingreso total hogar se desplaza desde el nivel más bajo (Rango 1), que se encuentra cercano a las carreras de la FCS, FCQS, hasta los más altos, que corresponden a las carreras de la FCA y FIC. Los valores medios - altos (Rangos 5, 7) están cercanos a la carrera de Medicina en la FCQS. Las carreras del área social son preferidas por estudiantes que provienen de familias con bajos ingresos, lo cual es congruente con la observación de que las becas de estudio, de la variable Origen recursos estudio, se han direccionado de manera preferente a estudiantes de la FCS.

Por otra parte, se observa que la mayoría de los estudiantes de la universidad reside en zonas urbanas, pero la categoría Zona rural se acerca más a las carreras de la FCS. Además, los estudiantes de la FCS y FCE provienen, mayoritariamente, de colegios fiscales y municipales; los niveles de formación académica de padres y madres de los estudiantes son más bajos en estos grupos, donde es usual encontrar casos de educación básica incompleta, formación en centros de alfabetización y ninguna formación. Los estudiantes que provienen de colegios particulares y fismisionales están con mayor frecuencia en las carreras de la FCQS, FCA y FIC; asimismo, los niveles más altos de formación de los padres y madres, como Postgrado Ph. D, Maestrías y Especializaciones médicas se asocian a carreras como Medicina e Ingeniería Civil. El análisis de las variables que se manifiestan con mayor presencia en la dimensión 2 del MCA indica que los grupos de estudiantes más jóvenes prefieren carreras relacionadas con las ciencias médicas y de la salud, ciencias naturales y exactas, ingenierías y tecnologías y ciencias agrícolas; mientras que, los grupos de mayor edad se asocian a carreras que se ubican en el área de las ciencias sociales y las humanidades. La FCA y FIC reportan menor frecuencia de casos de estudiantes con discapacidades que las demás facultades.

#### 4.2.2. Gráfico de control multivariante T2 Hotelling



**Figure 12.** Gráfico de control T2 Hotelling aplicado a las tablas de los periodos académicos analizados.

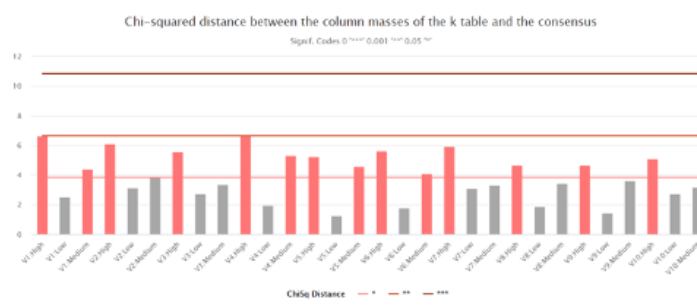
La figura 12 muestra el gráfico de control T2 de Hotelling para la representación de las  $k = 4$  tablas analizadas, éstas se representan por los puntos del gráfico y corresponden a los cuatro periodos académicos considerados en este estudio. El punto que representa al periodo académico 2019-1 ha sido reportado como un valor fuera de control, en consecuencia, será necesario un análisis de sus datos comparados con los de la tabla consenso (Figura 11) para identificar las causas de la variación y, si fuera el caso, tomar las acciones pertinentes. Para ello se realiza un MCA a la tabla 2019-1.





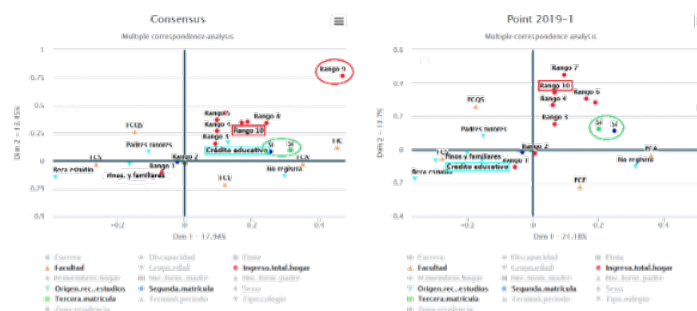
**Figure 13.** Comparación de los gráficos del MCA aplicado a la tabla consenso y la tabla 2019-2.

La figura 14 contiene los gráficos del MCA de la tabla consenso y la tabla 2019-1. Los puntos allí registrados corresponden a las 16 variables cualitativas con sus respectivas categorías. Se ve que hay puntos que conservan su ubicación o que han variado muy poco en ambas tablas, como las facultades, carreras, la variable Sexo y la Zona residencia. Además, se observa categorías de variables que han cambiado su ubicación de manera sensible y que pueden estar ocasionando el estado fuera de control. La identificación de estas tablas se facilita cuando se analiza la figura 14.



**Figure 14.** Distancia  $\chi^2$  entre las masas de la tabla consenso y las k tablas, Estudiantes 2019 2020.

La figura 14, permite apreciar, en un gráfico de barras, la distancia Chi cuadrado entre las categorías de la tabla consenso y de la tabla 2019-1, reportada como fuera de control. Mientras más altas son las barras, mayor es esta distancia. Las barras que sobresalen representan a las variables que, en la comparación, tienen una distribución muy distinta, de manera que han alcanzado significancia estadística muy alta y p-valores inferiores a 0.001 (tres asteriscos), por eso han adoptado el color correspondiente a ese nivel en el gráfico. Estas variables son las que con mayor fuerza están provocando el desplazamiento de la media del proceso y llevando al punto a un estado fuera de control. En consecuencia, es en ellas que se debe profundizar el análisis comparativo mediante el MCA.



**Figure 15.** Comparación de los gráficos del MCA con énfasis en las variables de mayor significancia estadística.

En la figura 15, la barra más alta representa a la categoría Rango 9 de la variable Ingreso total hogar, que se ubica en la esquina superior del cuadrante 1 en la tabla consenso, pero, ya no aparece en la tabla 2019-1. Las categorías Rango 8 y Rango 10 tuvieron un desplazamiento hacia la izquierda. Durante el periodo de estudio, los estudiantes que provienen de familias con ingresos más altos han ido migrando desde Ingeniería Civil, Tecnologías de la Información, Acuicultura, Medicina Veterinaria y Agronomía, hacia carreras como Medicina e Ingeniería Química; los de ingresos medios se van alejando de las carreras como Administración de Empresas, Economía, Contabilidad y Auditoría, mientras que los estudiantes de bajos recursos no han modificado su preferencia: Carreras de Educación, Sociología y Trabajo Social.

Otra variable que demuestra alta incidencia en el desplazamiento de la media del proceso es Origen recursos estudios. Se observa que la categoría Crédito educativo demuestra un desplazamiento considerable, en la tabla consenso se ubica en el primer cuadrante, mientras que, en la tabla 2021-1 aparece en el tercero (figura 15). Esto implica que el crédito educativo que se ofrece a los estudiantes ha cambiado de dirección, desde áreas relacionadas con las ciencias sociales y humanísticas hasta áreas administrativas, ingenierías, tecnologías y ciencias agrícolas. Por otra parte, la categoría Hermanos y familiares se mantiene cerca de las carreras de la FCS, lo que significa que los estudiantes de bajos recursos que estudian carreras de áreas sociales y humanísticas obtienen ayuda económica de sus familiares.

El nivel Postgrado Ph.D. de las variables Nivel de formación del padre y Nivel de formación de la madre manifiesta diferencias altamente significativas en los dos gráficos de MCA, pues sólo aparece en la tabla consenso, no en la 2019-1. Esto se entiende porque a principios de 2019 todavía no había padres de familia de la UTMACH que ostentaran ese grado académico, y en la comunidad general eran pocos. No obstante, con el transcurso del tiempo varios de ellos, que estaban en proceso de formación de doctorado, han logrado titularse, además, otros padres de familia que ya tenían tal grado académico han matriculado a sus hijos en esta universidad.

Al hacer un análisis del comportamiento de la variable Etnia en los años 2019 - 2020, llama la atención que estudiantes que se autodefinieron como Negros, Afroecuatorianos y Mulatos, se alejan de carreras de áreas sociales y humanísticas para acercarse a otras del área administrativa, como Administración de Empresas, Contabilidad y Auditoría, Comercio internacional, Turismo, Mercadotecnia y Economía. Mientras que, estudiantes que se autodenominaron Indígenas, migraron desde éstas hacia otras carreras en áreas sociales y humanísticas.

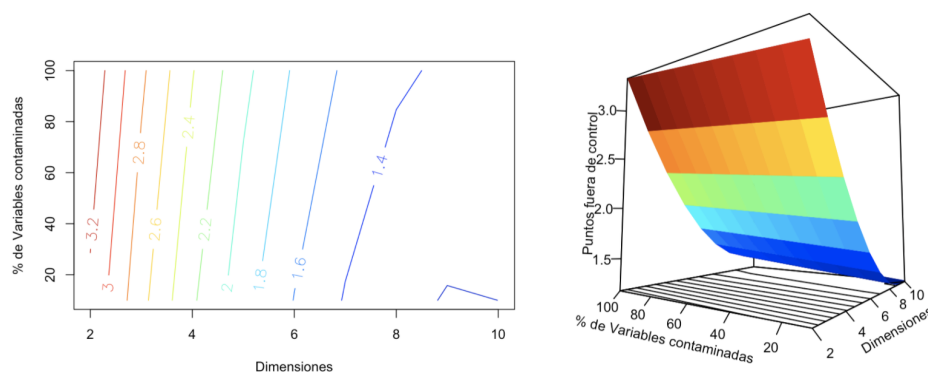
Como cambios relevantes en torno a la variable Discapacidad se tiene que, en los periodos académicos 2019-1 hasta 2020-2, la discapacidad Intelectual se acerca a la carrera de Derecho; la discapacidad Auditiva, a la Ingeniería química, Alimentos y Medicina. Mientras tanto, disminuye la frecuencia de estudiantes con discapacidad Visual en Gestión Ambiental, Artes plásticas y Pedagogía de la Actividad Física y Deporte.

## 5. Análisis de sensibilidad

Como se ha mencionado, en el gráfico T2Qv un punto fuera de control se interpreta como una tabla ( $k_i$ ) que incluye una cantidad o una proporción de variables contaminadas, de tal manera que la diferencia de los valores de masas de columna, entre de la matriz  $k_i$  y la matriz consenso, sean significativos según el valor  $p$  obtenido de la distribución  $\chi^2$ . En estos casos, se espera que los puntos en el gráfico T2Qv generalicen el comportamiento de estas diferencias y superen el límite de control superior ( $UCL$ ). La ubicación de este límite de control varía en función del número de dimensiones que se representen, así, cuando es alto se logra un desempeño óptimo, mientras que, se introduce inestabilidad y se pierde confiabilidad en los resultados al disminuir el número de dimensiones de entre las que se puede representar.

El gráfico de control propuesto es capaz de detectar un punto fuera de control, aún con un bajo número de variables contaminadas, cuando se trabaja con un alto número de dimensiones. Se recomienda  $p - 1$ , tal que  $p$  es el número total de dimensiones de la matriz inicial (Tabla 3). Cuando se

disminuye el número de dimensiones también disminuye la altura del límite de control superior (UCL), en consecuencia, se incrementa el número de puntos fuera de control, aunque no necesariamente las variables expresen diferencias significativas en sus valores, crece la probabilidad de falsos positivos. Por consiguiente, la pregunta que surge es hasta cuántas dimensiones se puede disminuir en el análisis sin perder confiabilidad en el resultado. La importancia de esta pregunta radica en la necesidad de disponer un gráfico confiable, que identifique puntos fuera de control aún si se ha aplicado a los datos una técnica de una reducción de dimensiones, sin caer en casos de falso positivo.



**Figure 16.** Curvas de nivel y superficie de respuesta obtenidas con el gráfico T2 Hotelling para variables cualitativas.

El análisis de sensibilidad utiliza curvas de nivel y superficies de respuesta (figura 16) para representar el número de puntos fuera de control, considerando el porcentaje de variables contaminadas de la  $k_i$  tabla y el número de dimensiones representadas. Los datos de prueba utilizados en el modelo se registran en 10 tablas, cada una de ellas incluye 10 variables y cada variable tiene tres categorías: alto, medio y bajo. La tabla 10 tiene una distribución diferente de las demás, esta es la tabla contaminada.

Se observa que el modelo es capaz de identificar un punto fuera de control trabajando con 9 dimensiones ( $p-1$ ), aún con un porcentaje bajo de variables contaminadas. Cuando el número de dimensiones disminuye a 8 y el porcentaje de variables contaminadas es cercano a 100%, detecta correctamente 1 punto fuera de control. Se observa además que cuando el número de dimensiones es menor se pierde estabilidad. En consecuencia, el análisis de sensibilidad ratifica que el gráfico de control T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones.

## Appendix A

### Appendix A.1

## Appendix B

## References

- Gutiérrez, H.; de la Vara Salazar, R. *Control estadístico de la calidad y seis sigma*; Vol. 3, McGraw Hill Education, 2013; p. 152 – 253.
- Ramos, M. Una alternativa a los métodos clásicos de control de procesos basada en coordenadas paralelas, métodos Biplot y Statis. PhD thesis, 2017.
- Li, J.; Tsung, F.; Zou, C. Directional control schemes for multivariate categorical processes. *Journal of Quality Technology* **2012**, *44*, 136–154.
- Hotelling, H. Multivariate quality control. *Techniques of statistical analysis*. McGraw-Hill, New York **1947**.

5. Lowry, C.A.; Woodall, W.H.; Champ, C.W.; Rigdon, S.E. A multivariate exponentially weighted moving average control chart. *Technometrics* **1992**, *34*, 46–53.
6. Crosier, R.B. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics* **1988**, *30*, 291–303.
7. APARISI, F. Hotelling's T<sup>2</sup> control chart with adaptive sample sizes. *International Journal of Production Research* **1996**, *34*, 2853–2862, [<https://doi.org/10.1080/00207549608905062>]. doi:10.1080/00207549608905062.
8. Aparisi, F.; Haro, C.L. Hotelling's T<sup>2</sup> control chart with variable sampling intervals. *International Journal of Production Research* **2001**, *39*, 3127–3140, [<https://doi.org/10.1080/00207540110054597>]. doi:10.1080/00207540110054597.
9. Faraz, A.; Parsian, A. Hotelling's T<sup>2</sup> control chart with double warning lines. *Statistical Papers* **2006**, *47*, 569–593. doi:10.1007/s00362-006-0307-x.
10. Shabbak, A.; Midi, H. An improvement of the hotelling statistic in monitoring multivariate quality characteristics. *Mathematical Problems in Engineering* **2012**, *2012*.
11. Kim, S.B.; Jitpitaklert, W.; Park, S.K.; Hwang, S.J. Data mining model-based control charts for multivariate and autocorrelated processes. *Expert Systems with Applications* **2012**, *39*, 2073–2081.
12. Ruiz-Barzola, O. Gráficos de Control de Calidad Multivariantes con Dimension Variable. PhD thesis, Universitat Politècnica de València, 2013.
13. Yeong, W.C.; Khoo, M.B.C.; Teoh, W.L.; Castagliola, P. A control chart for the multivariate coefficient of variation. *Quality and Reliability Engineering International* **2016**, *32*, 1213–1225.
14. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Systems with Applications* **2014**, *41*, 1701–1707. doi:10.1016/j.eswa.2013.08.068.
15. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Multivariate control chart based on PCA mix for variable and attribute quality characteristics. *Production & Manufacturing Research* **2018**, *6*, 364–384, [<https://doi.org/10.1080/21693277.2018.1517055>]. doi:10.1080/21693277.2018.1517055.
16. Liu, Y.; Liu, Y.; Jung, U. Nonparametric multivariate control chart based on density-sensitive novelty weight for non-normal processes. *Quality Technology & Quantitative Management* **2020**, *17*, 203–215.
17. YILMAZ, H.; Yanik, S. Design of Demerit Control Charts with Fuzzy c-Means Clustering and an Application in Textile Sector. *Textile and Apparel* **2020**, *30*, 117–125.
18. Farokhnia, M.; Niaki, S.T.A. Principal component analysis-based control charts using support vector machines for multivariate non-normal distributions. *Communications in Statistics - Simulation and Computation* **2020**, *49*, 1815–1838, [<https://doi.org/10.1080/03610918.2018.1506032>]. doi:10.1080/03610918.2018.1506032.
19. Xue, L.; Qiu, P. A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology* **2020**, pp. 1–14.
20. Ahsan, M.; Mashuri, M.; Wibawati.; Khusna, H.; Lee, M.H. Multivariate Control Chart Based on Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics. *Symmetry* **2020**, *12*. doi:10.3390/sym12111838.
21. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Outlier detection using PCA mix based T<sup>2</sup> control chart for continuous and categorical data. *Communications in Statistics - Simulation and Computation* **2021**, *50*, 1496–1523, [<https://doi.org/10.1080/03610918.2019.1586921>]. doi:10.1080/03610918.2019.1586921.
22. Holgate, P. Estimation for the bivariate Poisson distribution. *Biometrika* **1964**, *51*, 241–287.
23. Chiu, J.E.; Kuo, T.I. Attribute control chart for multivariate Poisson distribution. *Communications in Statistics-Theory and Methods* **2007**, *37*, 146–158.
24. Lee, L.H.; Costa, A.F.B. Control charts for individual observations of a bivariate Poisson process. *The International Journal of Advanced Manufacturing Technology* **2009**, *43*, 744–755.
25. Laungrungrong, B.; M, C.B.; Montgomery, D.C. EWMA control charts for multivariate Poisson-distributed data. *International Journal of Quality Engineering and Technology* **2011**, *2*, 185–211.
26. Epprecht, E.K.; Aparisi, F.; García-Bustos, S. Optimal linear combination of Poisson variables for multivariate statistical process control. *Computers & operations research* **2013**, *40*, 3021–3032.
27. Lu, X. Control chart for multivariate attribute processes. *International Journal of Production Research* **1998**, *36*, 3477–3489.

28. Mukhopadhyay, A.R. Multivariate attribute control chart using Mahalanobis D 2 statistic. *Journal of Applied Statistics* **2008**, *35*, 421–429.
29. Taleb, H.; Limam, M.; Hirota, K. Multivariate fuzzy multinomial control charts. *Quality Technology & Quantitative Management* **2006**, *3*, 437–453.
30. Taleb, H. Control charts applications for multivariate attribute processes. *Computers & Industrial Engineering* **2009**, *56*, 399–410.
31. Fernández, M.N.P.; García, A.C.; Barzola, O.R. Multivariate multinomial T 2 control chart using fuzzy approach. *International Journal of Production Research* **2015**, *53*, 2225–2238.
32. López, C.P. *Técnicas de análisis multivariante de datos*; Pearson Educación, 2004.
33. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series* **1901**, *2*, 417 – 441. doi:10.1080/14786440109462720.
34. Hotelling, H. Analysis of a complex of statistical variables into principal components. **1933**. *24*, 417 – 441. doi:10.1037/h0071325.
35. Ch, S.; others. General intelligence objectively determined and measured. *American Journal of Psychology* **1904**, *15*, 201–293.
36. Thurstone, L.L. Multiple-factor analysis; a development and expansion of The Vectors of Mind. **1947**.
37. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
38. Benzecri., J. OL'analyse des correspondances. En *L'Analyse des Données: Leçons sur L'analyse Factorielle et la Reconnaissance des Formes et Travaux*; Paris - 1973, 1973.
39. Michailidis, G.; Leeuw, J.D. The Gifi system of descriptive multivariate analysis. *Statistical Science* **1998**, pp. 307–336.
40. Gifi, A. *Nonlinear multivariate analysis*; Vol. 14, John Wiley & Sons, 1990.
41. López de Ipiña, F. Análisis multivriante aplicado al estudio del parentesco. Representaciones HOMALS. **2014**.
42. Torgerson, W.S. Multidimensional scaling: I. Theory and method. *Psychometrika* **1952**, *17*, 401–419.
43. Shepard, R.N. The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* **1962**, *27*, 125–140.
44. Escofier, B.; Pagès, J. Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis* **1994**, *18*, 121–140. doi:https://doi.org/10.1016/0167-9473(94)90135-X.